Ontology-based Feature Level Opinion Mining for Portuguese Reviews

Larissa A. de Freitas
PUCRS
FACIN
Porto Alegre, Brazil
larissa.freitas@acad.pucrs.br

Renata Vieira[*]
PUCRS
FACIN
Porto Alegre, Brazil
renata.vieira@pucrs.br

ABSTRACT

This paper presents a thesis whose goal is to propose and evaluate methods to identify polarity in Portuguese user generated reviews according to features described in domain ontologies (experiments will consider movie and hotel ontologies Movie Ontology[1] and Hontology[2]).

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms

Keywords

Feature Level, Ontology, Opinion Mining, Sentiment Analysis

1. PROBLEM

In the last decade humans have come to share their opinions in social media on the Web (e.g., forum discussions and posts in social network sites). Opinions are important because whenever we need to make a decision, we want to know others points of view. The increasing interest of industry and academia in opinion mining is partly due to its potential applications, such as: marketing, public relations and political campaign. It is important to mention that, although natural language processing have a long history, little research had been done about opinion text before 2000 [7].

Nowadays opinion mining has been investigated mainly in three level of granularity (document, sentence or feature). According to [6], both the document level and sentence level analyses do not discover what exactly people liked or not. Studying the opinion text, mainly feature level, is extremely challenging. For the ordinary user, it is too complex to analyse opinions about object and object features in the great number of online review sites and personal blogs on the Web.

In the literature, recent works about ontology-based feature level opinion mining are [14] and [8]. Our research intend to use ontology in the task of feature identification. In our work the term "features" or "object features" are represented by concepts, properties and individuals of ontologies – for example movie, isActorIn and "Ice Age 3", respectively. An object denote the target entity that has been commented on. Still, the orientation of an opinion on a feature indicates whether the opinion is positive, negative or neutral.

This paper is organized as follows: Section 2 presents state of the art; Section 3 describes the proposed approach; Section 4 details the methodology; preliminary results are discussed in Section 5; some conclusions and future work are finally presented in Section 6.

2. STATE OF THE ART

According to [2] opinion mining research has two main research directions: sentiment classification and feature level opinion mining [5][10]. The steps of feature level opinion mining are: to identify the object features in review, to decide whether the review is classified was positive, negative or neutral and, to summarize the discovered information.

Binali (2009) [3] presents an overview about feature level opinion mining, where: object extraction refers to the entity extraction in reviews (e.g., movie); objects features extraction refers the components (or parts) and attributes (or properties) extraction (e.g., title); object features sentiment detection refers to the sentiment about each feature (e.g., good title); object sentiment detection refers the global sentiment expressed in relation to an entity (e.g., recommended, not recommended); object comparing refers to the opinion between two entities (e.g., movie A and movie B); object features comparing refers to the opinion between two features entities (e.g., title movie A and title movie B).

Other recent works involving ontology and opinion mining have been proposed, such as: [14], [15], [8] and [11].

Ontologies provides a formal, structured knowledge representation, with the advantage of being reusable. They also provide a common vocabulary for a domain. In [8], the Movie Ontology describe in OWL[3] is enriched with WordNet[4] synonyms and populated with IMDb[5] informations (Figure 1). In [11], Formal Concept Analysis (FCA) is used for movie ontology construction. In [14] ontology construction is divided in two steps: at first sentences with con-

[*] advisor

[1] http://www.movieontology.org/2010/01/movieontology.owl

[2] http://ontolp.inf.pucrs.br/Recursos/Hontology/20120417.owl

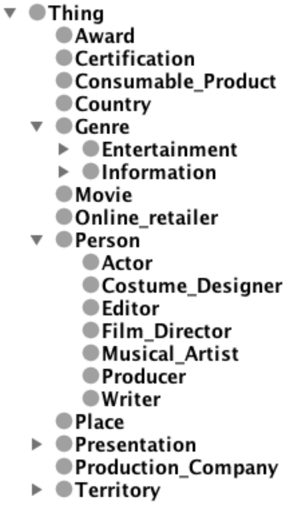
[3] http://www.w3.org/TR/owl-features/

[4] http://wordnetweb.princeton.edu/perl/webwn

[5] http://www.imdb.com/

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

junctions and seed concepts are selected; after the concepts are extracted. An hybrid model of ontology construction was adopted by [15]. The top-down process started with IMDb metadata, that is organized in five main sections (title, production status, cast, crew, and miscellaneous) and, the bottom-up process that involve content analysis.



- Thing
  - Award
  - Certification
  - Consumable_Product
  - Country
  - Genre
    - Entertainment
    - Information
  - Movie
  - Online_retailer
  - Person
    - Actor
    - Costume_Designer
    - Editor
    - Film_Director
    - Musical_Artist
    - Producer
    - Writer
  - Place
  - Presentation
  - Production_Company
  - Territory

Figure 1: Some concepts of Movie Ontology.

The literature shows that there are different levels of knowledge representation: authors using complex structures [8] [11] [15] – even if they not use all the knowledge available – and authors using just simple structures [14] for feature identification.

The only work that we found which reuse ontology is [8], the others build their own ontologies. A common point is the use of IMDb data.

In our work we intend to use complex structures and to use ontology in the feature identification step. Currently we are using only ontology classes. Unfortunately, the ontologies cited in [11], [14] and [15] are not available. The only ontology available that we found was the Movie Ontology. Our work focuses on obtaining an ontology-based feature level opinion mining method in the same line of [14] and [8], but considering Portuguese language.

Ontologies provide a structured knowledge representation and a common vocabulary for a domain (e.g. movie and hotel domain). Since the same opinion text can be presented in different levels of knowledge, ontology based opinion mining may help both for ordinary and expert users. For example, at the high level an opinion may refer to a film "genre" wheres at the low level the same opinion may be based on the concept "action". Besides, we intend develop linguistic rules for Portuguese, since linguistic rules vary from language to language.

3. PROPOSED APPROACH

This section presents the proposed approach (Figure 2). Similar to [14] and [8] approaches, our work is composed of four main steps. Initially, the algorithm receives as input a set of reviews which are pre-processed. After, explicit aspects are identified in the reviews using ontology terminology. The polarity measurement relies on a lexicon of tagged positive, negative, and neutral opinion words. Finally, in opinion mining module tuples with object features and polarity are generated.

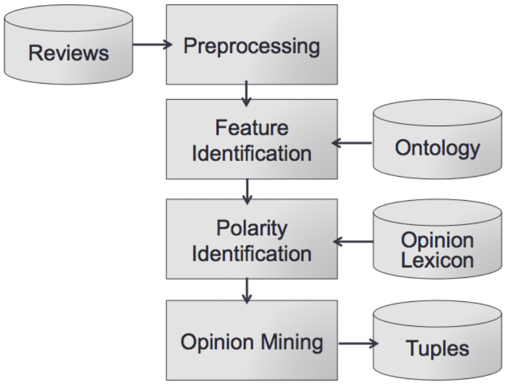


Figure 2: Overview of the method.

3.1 Preprocessing

A set of Portuguese reviews are pre-processed, in which we use tokenisers, sentence splitters, POS taggers and lemmatizers.

3.2 Feature Identification

After the pre-processing step, we use ontology (concepts, properties, instances and hierarchies) for feature identification. In [14] ontology concepts are identified in sentences. In [8] besides ontology concepts, properties and instances are identified in sentences. Differently of traditional feature identification methods, [8] calculates the feature importance based on frequency and position in the text. We intend to use the same strategy of [8], but instead of dividing reviews in three equal parts based on the number of words in the texts, we divide reviews in three equal parts based on the number of sentences in the texts.

3.3 Polarity Identification

For polarity identification we use list of adjectives, verbs, nouns and adverbs and their polarities. Portuguese opinion lexicons appeared in 2010 [12] [13]. SentiLex 2.0 [12] has 82.347 words and OpLexicon [13] has nearly 30.322 words in Portuguese. In the literature, many works about opinion mining use SentiWordNet[6]. SentiWordNet 3.0 [1] is a fragment of WordNet 3.0 manually annotated for positivity, negativity, and neutrality. Each synset has three numeric values (Pos, Neg e Obj) in the interval [0.0, 1.0]. Both [14] and [8] calculated polarity of the feature using SentiWordNet. This resource has nearly 117.000 words in English.

3.4 Opinion Mining

The opinion mining step generates a set of tuples containing features and their polarities. In our approach we use linguistic rules in windows of words.

4. METHODOLOGY

In this work the main idea is to apply ontology-based feature level opinion mining in Portuguese reviews. In reviews preprocessing we used the Portuguese TreeTagger[7]. The

[6] http://sentiwordnet.isti.cnr.it/

[7] http://gramatica.usc.es/ gamallo/tagger.htm

TreeTagger is a tool for annotating text with part-of-speech and lemma information. For example, for the sentence "Um dos melhores filmes que já vi!" ["One of the best movies I have ever seen!"] obtains the following lemmatized words accompanied by their grammatical categories (Table 1).

Table 1: Portuguese TreeTagger output.

Token	Tag	Lemma
Um	DET	um
dos	PREP+DET	de
melhores	ADJ	melhor
filmes	NOM	filme
que	PREP	que
já	V	<unknown>
vi	V	ver
!	SENT	!

4.1 Review Databases

For our studies we are considering two main review databases in the domain of movies (a common domain in this area) and the hotel review domain.

In our initial tests, we took 180 accommodation Portuguese reviews (10.706 words). The reviews are about Small and Medium Hotels in the Lisbon area, [4]. The information sources are Tripadvisor[8] and Booking.com[9].

The concepts used to identify explicit aspects in accommodation reviews are provided by Hontology. Its current version, has 282 concepts, 8 object properties and 31 data properties. Hontology is the first multilingual (Portuguese, English, Spanish and French) ontology for the accommodation domain.

And other database available contains 150 Portuguese movie reviews (8.999 words) extracted from the Omelete[10]. The concepts used to identify explicit aspects in movie reviews are provided by Movie Ontology. This ontology has 78 concepts, 30 object properties and 4 data properties.

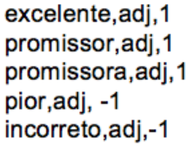
4.2 Opinion Lexicon

OpLexicon is a Opinion Lexicon provided by [13]. We removed of OpLexicon the list of adjectives in Portuguese contains 23.433 entries, such as excellent, promising, worse, and incorrect (Figure 3). This list is composed by the name of the adjective and a polarity which can assign one of three values: 1, -1 and 0.

Using the list of adjectives we analyse the opinion orientation expressed on each feature. If the word in a pre-processed review is a concept of ontology, we create a window of seven words and we search these words in the list of adjectives. At least one word of the window should be in the list of adjectives for the sentence polarity to be calculated.

4.3 Tuple Generation

For example, for the sentence "EXCELENTE!! O Ryan Gosling é o ator mais promissor da nova geração!! O cara consegue comover a gente até contracenando com uma boneca!! 5 OVOS!" ["EXCELLENT!! The actor Ryan Gosling is the most promising of the new generation!! The guy can move up we opposite with a doll! 5 EGGS!"] obtains the following window for actor concept "Gosling é o **ator** mais promissor da". We get 3 words before actor concept and 3 words after this concept. In the list of adjectives the word "promissor" [promising] have polarity 1 (positive). In summary the tuple generate in this case is (actor, positive).



excelente,adj,1
promissor,adj,1
promissora,adj,1
pior,adj, -1
incorreto,adj,-1

Figure 3: OpLexicon Example.

5. PRELIMINARY TESTS

We conducted some initial tests of proposed approach. The 180 accommodation comments and the 150 movie comments, which constitutes the sets used in our experiments were evaluated by a human. Polarity was classified as positive or negative. We evaluated the algorithm against the human evaluation.

The concepts hotel, room, and lunch were the most mentioned in the accommodation reviews. And the concepts movie and genre were the most mentioned in the movie reviews. Other concepts were not explicitly mentioned in the reviews therefore they are difficult to be detect polarity automatically.

According to these results, applying the algorithm in accommodation comments we find a recall for hotel concepts with positive polarity of 0.58 and precision of 0.10 [4]. For movie comments we find a f-measure for movie concepts with positive polarity of 0.62.

6. CONCLUSIONS AND FUTURE WORK

This paper introduced state-of-the-art and preliminary results about ontology based feature level opinion mining. So far no works about this topic was found for Portuguese language. The algorithm to deal with Portuguese online reviews is presented. Also, this work intends to summarize opinions about objects and objects features in different levels for the ordinary or expert users.

As future study, we will work on the improvement of the results reached by the algorithm. We plan to verify the inter-annotator agreement using the kappa coefficient. Moreover, we intend to use lemmatizer in preprocessing and properties, instances and hierarchies of ontologies in identification feature. Also, we plan to add list of verbs, list of adverbs and list of nouns in polarity identification. At last, we will apply a set of linguistic rules, such as: negatives, intensifiers [9] and irony/sarcasm detection. Linguistic rules, for example, vary from language to language.

Besides, we intend to study ways of solving problems such as different words (e.g., filminho and filmão) that refer to the same concept.

7. ACKNOWLEDGMENTS

We thank Aline Vanin for helping in review. We thank the Brazilian funding agency CAPES for the scholarship granted.

[8] www.tripadvisor.com

[9] www.booking.com

[10] omele.uol.com.br

8. REFERENCES

[1] A. Baccianella, S.; Esuli and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *7th International Conference on Language Resources and Evaluation*, 2010.

[2] T. Bhuiyan, Y. Xu, and A. Josang. State-of-the-art review on opinion mining from online customer's feedback. In *9th Asia-Pacific Complex Systems Conference*, 2009.

[3] H. Binali, V. Potdar, and C. Wu. A state of the art opinion mining and its application domains. In *International Conference on Industrial Technology*, 2009.

[4] M. Chaves, L. Freitas, M. Souza, and R. Vieira. Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector. In *17th International Conference on Applications of Natural Language Processing to Information Systems*, 2012.

[5] M. Hu and B. Liu. Mining opinion features in customer reviews. In *19th national conference on Artifical intelligence*, pages 755–760, 2004.

[6] B. Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, 2010.

[7] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 2008.

[8] I. Penalver-Martinez, R. Valencia-Garcia, and F. Garcia-Sanchez. Ontology-guided approach to feature-based opinion mining. In *International Conference on Applications of Natural Language to Information Systems*, 2011.

[9] L. Polanyi and A. Zaenen. Contextual valence shifters. *AAAI Spring Symposium on Attitude*, 20:1–10, 2004.

[10] A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346, 2005.

[11] K. P. P. Shein. Ontology based combined approach for sentiment classification. In *3rd International Conference on Communications and Information Technology*, CIT'09, pages 112–115, Stevens Point, Wisconsin, USA, 2009. World Scientific and Engineering Academy and Society.

[12] M. J. Silva, P. Carvalho, and L. Sarmento. Building a sentiment lexicon for social judgement mining. In *10th International Conference Computational Processing of the Portuguese Language*, 2012.

[13] M. Souza, R. Vieiras, D. Busetti, R. Chishman, and I. M. Alves. Construction of a portuguese opinion lexicon from multiple resources. In *8th Brazilian Symposium in Information and Human Language Technology*, 2012.

[14] L. Zhao and C. Li. Ontology based opinion mining for movie reviews. In *3rd International Conference Knowledge, Science, Engineering and Management*, 2009.

[15] L. Zhou and P. Chaovalit. Ontology-supported polarity mining. *Journal of the American Society for Information Science and Technology*, 59:98–110, 2008.